PREDICTION OF PM_{10} USING MULTIPLE LINEAR REGRESSION AND BOOSTED REGRESSION TREES

NUR HAZIQAH BINTI MOHD HAMID

SCHOOL OF CIVIL ENGINEERING
UNIVERSITI SAINS MALAYSIA
2017

PREDICTION OF PM_{10} USING MULTIPLE LINEAR REGRESSION AND BOOSTED REGRESSION TREES

By

NUR HAZIQAH BINTI MOHD HAMID

This dissertation is submitted to

UNIVERSITI SAINS MALAYSIA

As partial fulfilment of requirement for the degree of

BACHELOR OF ENGINEERING (HONS.)
(CIVIL ENGINEERING)

School of Civil Engineering,
Universiti Sains Malaysia

June 2017

USM

UNIVERSITI SAINS MALAYSIA

SCHOOL OF CIVIL ENGINEERING
ACADEMIC SESSION 2016/2017

FINAL YEAR PROJECT EAA492/6
DISSERTATION ENDORSEMENT FORM

Title: Prediction of PM_{10} using Multiple Linear Regression and Boosted Regression Trees

Name of Student: Nur Haziqah binti Mohd Hamid

I hereby declare that all corrections and comments made by the supervisor(s) and examiner have been taken into consideration and rectified accordingly.

Signature:

____

Date :

Approved by:

____

(Signature of Supervisor)

Name of Supervisor : Prof. Ahmad Shukri bin Yahaya

Date :

Approved by:

____

(Signature of Examiner)

Name of Examiner : Prof. Dr. Nor Azam bin Ramli

Date :

PREDICTION OF PM_{10} USING MULTIPLE LINEAR REGRESSION AND BOOSTED REGRESSION TREES

By

NUR HAZIQAH BINTI MOHD HAMID

This dissertation is submitted to

UNIVERSITI SAINS MALAYSIA

As partial fulfilment of requirement for the degree of

BACHELOR OF ENGINEERING (HONS.)
(CIVIL ENGINEERING)

School of Civil Engineering,
Universiti Sains Malaysia

June 2017

SCHOOL OF CIVIL ENGINEERING
ACADEMIC SESSION 2016/2017

FINAL YEAR PROJECT EAA492/6
SAMPLE CD COVER

SCHOOL OF CIVIL ENGINEERING
ACADEMIC SESSION 2016/2017

Supervisor:
Prof. Ahmad Shukri Yahaya

PREDICTION OF PM_{10} USING MULTIPLE
LINEAR REGRESSION AND BOOSTED
REGRESSION TREES

Prepared by:
Nur Haziqah Mohd Hamid
126960

ACKNOWLEDGEMENT

Alhamdulillah and praise to Allah s.w.t for giving me the chance, knowledge and strength to surpass all the hard work that I have pass through during completing this dissertation. First of all, I would like to express my big appreciation to my supervisor Professor Ahmad Shukri Yahaya for giving me the opportunity to be under his supervision. I really appreciate for the guidance and advice throughout completing my final year project. Without his supervision and support, this research would be difficult for me to complete with success.

Besides that, I would like to express my deepest appreciation to all lecturers of School of Civil Engineering that who are involved in helping and giving me briefings and explanations in completing this dissertation. Furthermore, I would like to dedicate my thanks to the Department of Environment for providing the data needed to carry out this research.

Finally, I would like to thank my dear parents, Mohd Hamid bin Abu Bakar and Che Zaitun binti Ishak from the bottom of my heart for always supporting me throughout the journey of completing my undergraduate studies. My sincere thanks are also dedicated to my senior, Zul Azmi bin Mokhtar and Norazrin binti Ramli for giving their guidance towards me to complete this dissertation. Last but not least, special thanks also goes to all my friends for their help and always giving me a spirit to do well.

ABSTRAK

Zarah berdiameter aerodinamik kurang daripada 10µm (PM_{10}) adalah salah satu daripada udara yang boleh menjejaskan kesihatan manusia. Tujuan kajian ini adalah untuk meramal kepekatan zarah untuk hari esok (PM_{10D1}) dengan menggunakan Model Regresi Linear Berganda (MLR) dan Model Regresi Pokok Penggalak (BRT). Data min setiap hari digunakan dari 2013 hingga 2015 dibahagikan kepada data latihan (70%) dan data pengesahan (30%). Parameter yang mempengaruhi kepekatan PM_{10} untuk hari seterusnya adalah zarah terampai (PM_{10D0}), kelajuan angin (WS), suhu (T), kelembapan relatif (RH), sulfur dioksida (SO_2), nitrogen dioksida (NO_2), ozon (O_3) dan karbon monoksida (CO). Data min harian telah dipilih di empat stesen pemantauan iaitu Jerantut (stesen latar belakang), Nilai (kawasan perindustrian), Seberang Jaya (kawasan sub-bandar) dan Shah Alam (kawasan bandar). Keputusan yang diperolehi menunjukkan bahawa setesen Nilai merekodkan nilai kepekatan PM_{10} tertinggi berbanding stesen-stesen lain. Sumbangan utama pencemar udara di stesen Nilai adalah zarah terampai (PM_{10D0}), karbon monoksida, nitrogen dioksida dan ozon. Keputusan yang diperolehi menunjukkan bahawa model Regresi Linear Berganda (MLR) adalah model yang terbaik untuk meramal kepekatan PM_{10} hari seterusnya berbanding model Regresi Pokok Penggalak (BRT).

ABSTRACT

Particulate matter with an aerodynamic diameter less than 10µm (PM_{10}) is one of the pollutants that can adversely affect human health. The aims of this study is to predict particulate matter concentration for the next day (PM_{10D1}) by using Multiple Linear Regression (MLR) and Boosted Regression Trees (BRT) models. The daily mean data used from 2013 until 2015 is divided into training data (70%) and validation data (30%). The parameters that influence PM_{10} concentration for the next day are particulate matter (PM_{10D0}), wind speed (WS), temperature (T), relative humidity (RH), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and carbon monoxide (CO). Daily mean data were selected at four monitoring stations which are Jerantut (background station), Nilai (industrial area), Seberang Jaya (sub-urban area) and Shah Alam (urban area). The results obtained shows that Nilai station recorded the highest mean value of PM_{10} concentration compared to other stations. The main contributions of air pollution at Nilai station are particulate matter (PM_{10D0}), carbon monoxide, nitrogen dioxide and ozone. The result shows that Multiple Linear Regression models (MLR) is the better model to predict the next day of PM_{10} concentration compared to Boosted Regression Trees (BRT).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

API	Air Pollution Index
BAM	Beta Attenuation Monitor
BRT	Boosted Regression Trees
CO	Carbon monoxide
DoE	Department of Environment
MAAQG	Malaysian Ambient Air Quality Guidelines
MAAQS	Malaysian Ambient Air Quality Standard
MLR	Multiple Linear Regression
NO_2	Nitrogen dioxide
O_3	Ozone
PM_{10}	Particulate matter with an aerodynamic diameter less than 10μm
RH	Relative humidity
SO_2	Sulphur dioxide
SPSS	Statistical Package and Services Solution
T	Temperature
USEPA	United States Environmental Protection Agency
WS	Wind Speed

CHAPTER 1

INTRODUCTION

1.1 Background

Sansuddin et al., (2011) stated that air pollution have become common phenomena in urban and industrial areas in Malaysia due to increasing quantities of pollutants emitted into the atmosphere by local anthropogenic sources (human activity). Anthropogenic air pollution from sources like motor vehicles and industries caused serious harm to human health at highly populated urban areas.

PM_{10} is particulate matter with an aerodynamic diameter less than 10μm, which can be a suspension of solid, liquid or a combination of solid and liquid particles in the air. It is one of the main pollutants that can adversely affect human health. Sustained exposure to PM_{10} has long been associated with effects on the respiratory system, damage to lung tissue, cancer, and premature death (Ul-Saufie et al., 2013).

According to Department of Environment Malaysia, PM_{10} are emitted from motor vehicle exhausts, heat and power generation plants, industrial process and open burning activities. However, the most prominent sources are industry and heavy traffic (Ul-Saufie et al., 2012).

The air quality status in Malaysia are based on the air quality monitoring in several cities for the hourly concentration of carbon monoxide (CO), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and particulate matter less than or equal to 10μm (PM_{10}) (Ul-Saufie et al., 2013). The status of air quality can be divided into five categories which are good, moderate, unhealthy, very unhealthy and hazardous as presented in Table 1.1.

Table 1.1 : Malaysia Air Pollution Index (API)
(Source : Department of Environment, 2017)

API	Status	Health Effect	Health Advice
0-50	Good	Low pollution without any bad effect on health.	No restriction for outdoor activities to the public. Maintain healthy lifestyle.
51-100	Moderate	Moderate pollution that does not pose any bad effect on health.	No restriction for outdoor activities to the public. Maintain healthy lifestyle.
101-200	Unhealthy	Worsen the health condition of high risk people who is the people with heart and lung complications.	Limited outdoor activities for the high risk people. Public need to reduce the extreme outdoor activities.
201-300	Very Unhealthy	Worsen the health condition and low tolerance of physical exercises to people with heart and lung complications. Affect public health.	Old and high risk people are prohibited for outdoor activities. Public are advised to prevent from outdoor activities.
>300	Hazardous	Hazardous to high risk people and public health.	Old and high risk people are prohibited for outdoor activities. Public are advised to prevent from outdoor activities.
>500	Emergency	Hazardous to high risk people and public health.	Public are advised to follow orders from National Security Council and always follow the announcement in mass media.

In Malaysia, Air Pollutant Index (API) has been used as an indicator of air quality since 1989. The calculation of the API value is described in the DoE (1997) report. The API value is based on the sub-index of five criteria of air pollutants (PM_{10}, SO_2, NO_2, CO and O_3). The highest sub-index value of the individual pollutants is taken as the API value for that particular period of time (Awang et al., 2000). To determine the API for a given time period, the sub-index values (for all five air pollutants included in the API system) were calculated based on the average concentration calculated. The maximum sub-index of all five pollutants was selected as the API and the specific responsible air pollutants for the API value has to be reported to indicate the relevant health effect category and actions to be taken. The process flowchart for calculating API value at a given time was detailed in Figure 1.1 below.

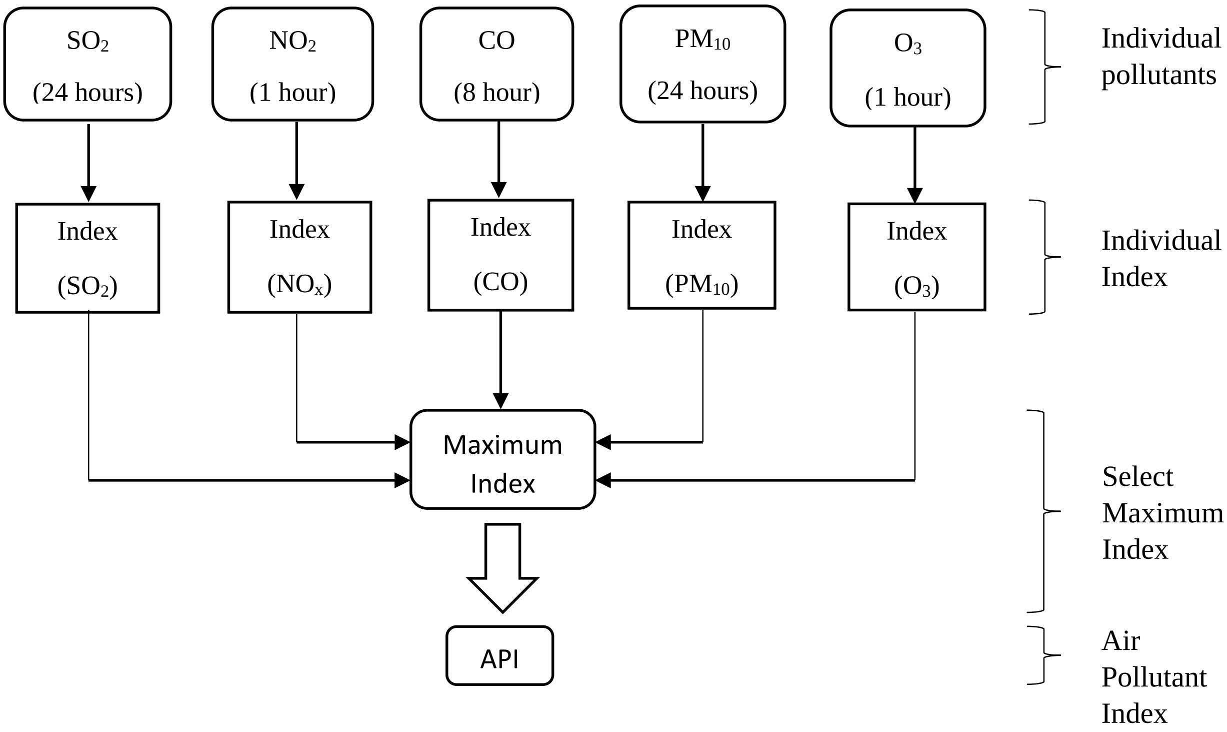

Figure 1.1 : API calculation (Source : Department of Environment, 2017)

As the main pollutant involved in this research, DoE has used new standard in the MAAQS Interim Target 1 (IT-1) in 2015 which state that the average threshold limit concentration for particulate matter (PM_{10}) is at 150 µg/m^3 for a 24-hours period and 50 µg/m^3 for annual. If the concentrations exceed the limit values, health problems may occur to the surrounding people. Malaysian Ambient Air Quality Standard (MAAQS) were issued and target values for annual and daily mean mass concentrations for various air pollutant were established to control and reduce air pollutant levels in the atmosphere. Table 1.2 shows the Malaysian Ambient Air Quality Standard (MAAQS).

Table 1.2 : Malaysian Ambient Air Quality Standard (MAAQS)
(Source : Department of Environment, 2017)

Pollutants	**Averaging Time**	**Ambient Air Quality Standard**		
		IT-1 (2015)	IT-2 (2018)	IT-3 (2020)
		(μg/m^3)	(μg/m^3)	(μg/m^3)
PM_{10}	1 Year 24 Hours	50 150	45 120	40 100
$PM_{2.5}$	1 Year 24 Hours	35 75	25 50	15 35
Sulphur Dioxide (SO_2)	1 Hour 24 Hours	350 105	300 90	250 80
Nitrogen Dioxide (NO_2)	1 Hour 24 Hours	320 75	300 75	280 70
Ground Level Ozone (O_3)	1 Hour 8 Hours	200 120	200 120	180 100
*Carbon Monoxide (CO)	1 Hours 8 Hours	35 10	35 10	30 10

Note : *mg/m^3

1.2 Problem Statement

The presence of PM_{10} in atmosphere can cause severe health impacts to human such as allergies, coughing, asthma, respiratory related illnesses, nose and throat irritations, premature mortality, irregular heartbeat and more severe impacts such as hospital admission. PM_{10} may also bring negative impacts on the growth and productivity of small and short cycle plant species such as vegetables (Ul-Saufie et al., 2012).

The main sources of air pollutants in Malaysia are mobile sources particularly motor vehicles exhausts, stationary and transboundary emission. Stationary sources include heat and power generation plants, industrial waste incinerators, the emission of dust from urban construction works and quarries, along with open burning activities. The transported air pollution from forest fires from neighbouring countries is referred to as transboundary pollution (Dominick et al., 2012).

Urban air pollution, with its long- and short-term impacts on human health, well-being and the environment, has been a widely recognised problem over the last 50 years The clear phenomenon of rural to urban migration has brought as a consequence greater emissions into the atmosphere, which have predominantly been produced by the increase in traffic. In addition, the rapid growth of urbanisation and industrialisation where the progressive expansion of suburbs into closer proximity with industrial plants in certain areas has led to the problem of air pollution becoming an increasingly important issue (Azmi et al., 2010).

According to Department of Occupational Safety and Health, (2017), outdoor activity involving strenuous activity should be minimised when API (Air Pollution Index) is more than 100 which means that air quality is deemed unhealthy and employees will be exposed to higher levels of safety and health risk due to poor visibility and/or ill effects of haze. Employers have a duty to protect their employees’ safety and health at working area. Employees must use respiratory protective devices (or mask) especially for those who are working outdoors. If they experience breathing difficulty from wearing respirators while working outdoors, employers should deploy them to work indoor where the pollutant concentration is lower. When API value exceed 200, all susceptible employees should be deployed to work indoor, preferably in work that is not physically strenuous. Furthermore, daily activities also will be affected due to unhealthy event.

According to Department of Environment (2017), during Southwest Monsoon, Malaysia had experienced deterioration of air quality from August to September 2015 due to massive land and forest fires in Sumatra and Kalimantan, Indonesia. On 15 September 2015, 34 areas in the country recorded unhealthy air quality status for the first time in Malaysia's history since 1997. Due to the API reading reaching to 200, all schools in Putrajaya, Kuala Lumpur, Selangor, Negeri Sembilan and Melaka were closed on 15 September 2015 while all schools in Kuching and Samarahan Divisions, Sarawak were closed on 18 September 2015. The highest API reading was 211 (very unhealthy), in Banting, Selangor on 14 September 2015. Hence, future PM_{10} concentration prediction is very important as it can help local authorities to enact preventative measures to reduce the impact of air pollution (Ul-Saufie, 2013). Other than that, local authorities are able to give an early warning to people who are in risk of acute and chronic health effects from air pollution.

1.3 Objectives

The main objectives of this study are :

1. To determine the characteristic of PM_{10}, weather parameters (relative humidity, temperature, wind speed) and gaseous parameters (SO_2, NO_2, CO and O_3).
2. To determine PM_{10} concentration models by using multiple linear regression and boosted regression trees.
3. To determine the best model to predict PM_{10} concentration.

1.4 Scope of Research

Four stations are selected for predicting PM_{10} concentration which are Seberang Jaya (sub-urban area), Shah Alam (urban area), Nilai (industrial area) and Jerantut which acts as a background station situated in central Peninsular Malaysia. Daily data obtained from Continuous Air Quality Monitoring (CAQM) stations is used from 2013 until 2015.

The parameters used are temperature (T; °C), relative humidity (RH; %), sulphur dioxide (SO_2; ppm), nitrogen dioxide (NO_2; ppm), ozone (O_3; ppm), particulate matter (PM_{10}; $\mu g/m^3$) and carbon monoxide (CO; ppm). Multiple linear regression and boosted regression tree model are used to determine the best prediction concentration for PM_{10} to develop the best model for predicting concentration of PM_{10}.

1.5 Thesis Outline

This thesis has five chapters and the brief outlines of this thesis are as follows :

Chapter 1 is an introduction about air pollution, particulate matter (PM_{10}) and the sources of air pollution in Malaysia. This chapter also discussed about the problem statement, objectives and scope of research.

Chapter 2 basically presents the literature reviews and related previous study of particulate matter (PM_{10}), sources of PM_{10} concentration, effects of PM_{10} concentration, air quality data and statistical modelling in predicting PM_{10} concentration.

Chapter 3 describes the methodology of the study. Procedures applied to predict PM_{10} concentration, method that were used and data collection of this research are also discussed in this chapter.

Chapter 4 discussed about the results obtained from the data analysis by using Statistical Package and Service Solution for multiple linear regression, R Studio for boosted regression trees and Matlab for performance indicators. All the data analysis are summarized in the table and performance indicator is obtained to determine the best model for each model.

Chapter 5 concludes about this study and provide some appropriate recommendations for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter discussed about the previous researches that have been done regarding the application of multiple linear regression and boosted regression trees for the purposes of modelling and prediction of air pollutant in the environment area. Furthermore, previous studies was also reviewed as an additional knowledge and guidance for this research.

2.2 Particulate Matter (PM)

Particular matter is the term given to the tiny particles of solid or semi-solid material found in the atmosphere. Particulate matter with an aerodynamic diameter of less than 10μm (PM_{10}) has been identified as an important atmospheric pollutant in major cities in Southeast Asia, particularly the Klang Valley, Malaysia. It is believed to have an effect on the human respiratory system which in turn may result in chronic obstructive pulmonary disease and asthma. In Malaysia, PM_{10} is one of the major air pollutants and is decisive in the computation of Malaysian Air Pollution Index (MAPI) (Juneng et al., 2011).

Epidemiological studies have found that atmospheric particulate matter, especially PM_{10} (inhalable particulate matter with aerodynamic diameter less than or equal to 10μm) is one of the pollutants that are harmful to human health (An et al., 2013).

2.3 Sources of PM_{10} concentration

Previous study shows that severe problems in air quality status in the Malaysian Peninsular only exist in highly urbanised areas. This is particularly true with respect to dust fall-out, suspended particulate matter and lead to the ambient air along congested roadsides. These problems are largely attributed to motor vehicles emissions (Awang et al., 2000).

PM_{10} is emitted by both natural and anthropogenic sources. Example of natural causes of air pollution are volcanic eruptions, forest fires and windblown dust while anthropogenic sources are like motor vehicles and industries (Jamal et al., 2004). Department of Environment (2010) listed five major PM_{10} emissions in Malaysia, these being motor vehicle exhaust, heat and power plants, industrial sources, and open burning. But, according to Ul-Saufie et al. (2012), the most prominent sources are industry and heavy traffic. The major sources of air pollution in Malaysia, notably in urban areas are motor vehicles (mobile sources), industrial and power plants (stationary sources) and open burning (Afroz et al. 2003).

2.3.1 Mobile sources

According to the Department of Statistics Malaysia (2015), the percentage of transport usage on the road is increasing year by year due to the population in Malaysia has increase annually. Anthropogenic air pollution from sources such as motor vehicles and industries continues to affect human health as it produce unhealthy events.

Located in the centre of South East Asia, Malaysia is experiencing rapid urban growth and is affected by local and regional air pollution (Latif et al., 2011).

Urbanisation, with industrial development has contributed to high amounts of atmospheric pollutants. Traffic is the major source of air pollution in highly urbanised areas in most developing countries on the Malaysian Peninsular (Azmi et al., 2010). This is true with respect to dust fall-out, suspended particulate matter and lead in the ambient air along congested roads. These problems are mainly attributed to motor vehicle emissions (Awang et al., 2000).

2.3.2 Stationary sources

Stationary sources include power plants, industrial waste incinerators, the emission of dust from urban construction works and quarries, along with open burning. The transported air pollution from forest fires from neighbouring countries is referred to as transboundary pollution (Othman et al., 2014).

According to Department of Environment (2017), in August 2005 haze episode was considered as more severe to the previous episode in 1997 as far as Peninsular Malaysia is concerned when the whole part of Klang Valley and its surrounding areas were badly affected by the haze events. It reached its peak when a Haze Emergency was declared on 11 August 2005 in two areas, namely Port Klang and Kuala Selangor as the Air Pollution Index (API) in both areas exceeded 500. The Haze Emergency was then lifted on 13 August 2005 after the API readings in both areas dropped below the hazardous level (301) and visibility improved.

2.3.3 Open burning

Open burning sources of air pollution in Malaysia include the burning of solid wastes and forest fires. This is common at some poorly managed disposal sites and results in smoke and fly ash problems. According to Department of Environment (2017), during the dry period between February and Mac 2014, Peninsular Malaysia had experienced moderate haze episode where air quality deteriorated to unhealthy and hazardous levels. The affected areas and states were the Klang Valley, Perak, Melaka, Negeri Sembilan and Johor. The haze was due to forest and peatland fires in several states namely in Selangor, Perak, Pahang, Johor, Kedah, Kelantan and Terengganu. The haze episode worsened on 14 March 2014 as the API level rose to hazardous level (API more than 300) in two areas namely Port Klang and Banting, Selangor. The haze situation had caused 203 schools in the Klang and Kuala Langat Districts in Selangor to be closed as the API reached very unhealthy levels of more than 200.

2.4 Effects of PM_{10} concentration

Urban air pollution, with its long and short-term impacts on human health, well-being and the environment, has been a widely recognised problem over the last 50 years. The clear phenomenon of rural to urban migration has brought as a consequence greater emissions into the atmosphere, which have predominantly been produced by the increase in traffic (Azmi et al., 2010).

2.4.1 Effects of PM_{10} concentration on human health

According to United States Environmental Protection Agency (USEPA), particulate matter (PM) is an air pollutant consisting of a mixture of solid and liquid particles suspended in the air with diameter less than or equal to 10μm (PM_{10}). It can cause significant health effects, particularly among the elderly and infants, people with asthma and other respiratory diseases. Some particles are released directly from a specific source, while others are form in complicated chemical reactions in the atmosphere. Particles come in a wide range of sizes as shown in Figure 2.1. Particles less than or equal to 10μm in diameter are so small that they can get into the lungs, potentially causing serious health problems since 10μm is less than the width of a single human hair.

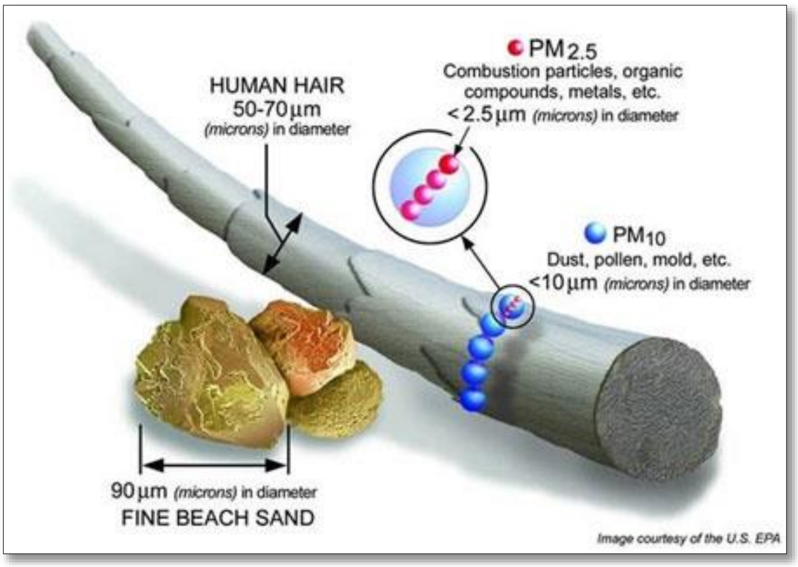
PM2.5
Combustion particles, organic compounds, metals, etc.
<2.5µm (microns) in diameter
HUMAN HAIR
50-70µm
(microns) in diameter
PM10
Dust, pollen, mold, etc.
<10µm (microns) in diameter
90µm (microns) in diameter
FINE BEACH SAND
Image courtesy of the U.S. EPA

Figure 2.1 : Size comparison of $PM_{2.5}$ and PM_{10}
(Source : United States Environmental Protection Agency, 2017)

Generally, the smaller a particle is, the more deeply it will penetrate to deposit on the respiratory tract at an increasing rate. In nasal-breathing, the cilia and the mucus act as a very effective filter for most particulates exceeding 10μm in diameter (coarse PM). Because the coarse PM fraction settles quickly, it tends to lodge in the trachea (upper throat) or in the bronchi. During inhale this PM will be initially collected in nose and throat. Then, the body will react to eliminate these intruding PM through such processes as sneezing and coughing (Kim et al., 2015).

Exposure to PM has been identified as the cause of numerous health effects including increased hospital admissions, emergency room visits, respiratory symptoms, exacerbation of chronic respiratory and cardiovascular diseases, decreased lung function, and premature mortality (Kim et al., 2015).

2.4.2 Effects of PM_{10} on environment

PM_{10} may affect animals in the same way as it affects humans. Particles in general, not specifically PM_{10} or $PM_{2.5}$, affect the aesthetics and utility of areas through visibility reduction and may affect buildings and vegetation. The specific effect of particles depends on their composition, concentration and the presence of other pollutants such as acid forming gases. Particles in the air affect both the quality of the air and visibility. Once in the air particulate matter generally takes a long time to settle. The particulates may be washed from the air by rain or snow. When they settle on land they may settle permanently. In water, particulates may settle, dissolve or both. PM_{10} and $PM_{2.5}$ are very fine and light and are therefore easily entrained into the air by wind or disturbances. Chemical changes may occur, as may reactions with other substances, depending on the composition of the particles. Particles may stick together or break apart, changing the size distribution over time (Department of Environment and Energy, Australia, 2013).

2.5 Air Quality Data

The air quality data used for this analysis were obtained from the Air Quality Division of the Department of the Environment, Malaysia, (DoE) through long-term monitoring by Department of Environment. The Department of Environment is one of the bodies in Malaysia that is responsible in monitoring the status of air quality throughout the country to perceive any significant change which may cause harm to human health and environment.

The PM_{10} concentration data used in this study was recorded as part of a Malaysian Continuous Air Quality Monitoring (CAQM) program, using the Beta Attenuation Mass Monitor (BAM-1020) as manufactured by Met One Instruments Inc. The monitoring network was installed, operated and maintained by Malaysian Department of the Environment (Afroz et al., 2003).

The air pollutants concentration are measured by 52 Continuous Air Quality Monitoring Stations (CAQMS) in Malaysia. This 52 monitoring station was categorized into four categories which is industrial, urban, sub-urban and background station. There are five major air pollutants and meteorological parameters used that are ground level ozone (O_3), carbon monoxide (CO), nitrogen oxide (NO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2), particulate matter with diameter size below than 10μm (PM_{10}), particulate matter with diameter size below 2.5μm ($PM_{2.5}$), ambient temperature, relative humidity and wind speed.

Figure 2.2 shows the beta attenuation monitor (BAM-1020) to measure and record hourly particulate mass concentration in ambient air by using beta ray attenuation to calculate collected particle mass concentration in unit of μg/m^3. BAM-1020 can measure the concentration in minutes, hours and days. At the beginning of the hour, a small ^{14}C (carbon-14) element emits beta rays through a clean spot of filter tape to determine a zero reading. These beta rays are detected and counted by a sensitive scintillation detector to determine a zero reading. The BAM-1020 then advances this spot of tape or exact spot to the sample nozzle, where air containing particulate is sample onto the filter tape. At the end of the hour, the same exact dirty spot is placed back between the beta source and the detector, thereby causing an attenuation of the beta ray signal which is used to determine the mass of the particulate matter on the filter tape. Then, the mass is used to calculate the volumetric concentration of particulate matter in ambient air (Met One Instrument, 2017).

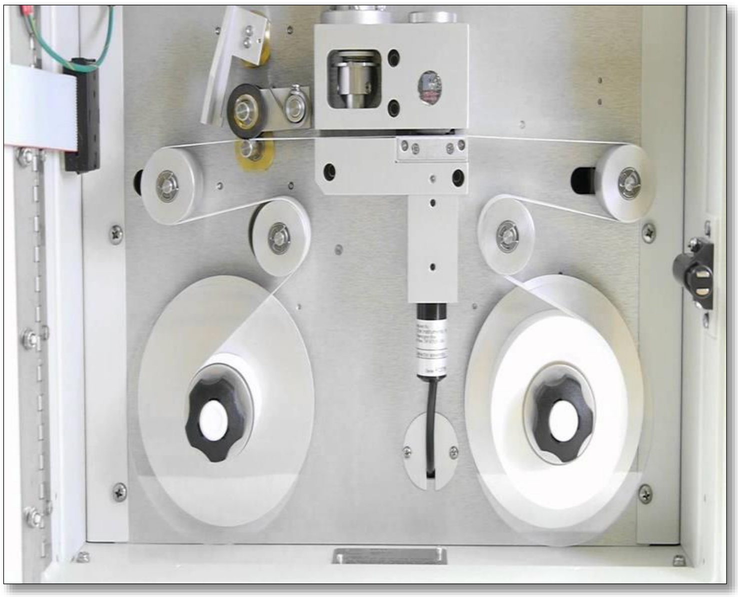

Figure 2.2 : Beta Attenuation Monitor (BAM-1020)
(Source : California Environmental Protection Agency, 2017)

2.6 Parameters

Parameters that were used in this research are particulate matter less than 10μm ($PM_{10,}$ μg/m^3), temperature (T; °C), relative humidity (RH; %), sulphur dioxide (SO_2; ppm), nitrogen dioxide (NO_2; ppm), ozone ($O_{3;}$ ppm) and carbon monoxide (CO; ppm). Air Pollutant Index (API) is an indicator for the air quality status at any particular area. The API value is calculated based on average concentration of air pollutants. The air pollutant with the highest concentration (dominant pollutant) will determine the API value.

2.6.1 Sulphur Dioxide (SO_2)

A study by Brunelli et al., 2007 defined that SO_2 is the natural product of sulphur. It is a colourless gas, with a prickly odour. SO_2 is released from those combustion processes that use fossil fuels (gas-oil, combustible oil, coal) where sulphur is present as an impurity, and from metallurgical tests.

2.6.2 Nitrogen Dioxide (NO_2)

Nitrogen dioxide is a toxic gas with a strong and prickly odour and with great irritating power; it is a very reactive, highly corrosive, energetic oxidizer. The well known yellowish colour of the hazes that cover the cities is due to the nitrogen dioxide caused by elevated vehicular traffic. It represents a secondary pollutant because it derives from nitrogen monoxide oxidation in the atmosphere. NO_2 mainly acts as an oxidant able to damage membranes and cellular proteins. Nitrogen oxide finally, contributes to the formation of acid rain, favouring the accumulation of nitrates in the ground and in water, can cause alterations of the environmental (eutrophication) ecological equilibriums (Brunelli et al., 2007).

2.6.3 Ozone (O_3)

Ozone is a toxic gas of a bluish colour, made of unstable molecules formed by three oxygen atoms (O_3). Due to these characteristics ozone is an energetic oxidant able to demolish both organic and inorganic material. In the troposphere it is present in low concentrations and it represents a particularly insidious secondary pollutant. It is produced during various chemical reactions in the presence of sunlight starting from the primary pollutants, in particular from the nitrogen dioxide (Brunelli et al., 2007).

2.6.4 Carbon Monoxide (CO)

The most important source of CO in the urban areas is represented by vehicular traffic. The CO natural background level varies between 0.01 and 0.23 mg/m^3. Especially CO concentrations in micro-environments can reach extremely elevated levels (even over 115 mg/m^3 for different times) for insufficient ventilation. CO concentrations (up to 60 mg/m^3) are also found inside domestic environments, due to the use of gas heaters. The toxic effects of acute exposure to carbon monoxide are due to the ability of the CO to bind with blood haemoglobin, forming the COHb, thus reducing the blood ability to transport oxygen into the various parts of the body (Brunelli et al., 2007).

2.7 Statistical Analysis in Modelling and Predicting Particulate Matter

Statistical modelling could offer good insights in short term predicting of future air pollution levels (next day, next 2-day and next 3-day), hence allowing local environmental authorities to carry out daily air pollution forecasts. The public health advisors can use this information to make decisions regarding air pollution abatement measures (Ul-Saufie et al., 2012).

2.7.1 Multiple Linear Regression (MLR)

Regression techniques had been used for a long time as forecasting tools in many fields, especially in air pollution forecasting. It determines the linear relationship between selected parameters, which the models established might be less accurate in forecasting complex situation. Multiple linear regression (MLR) is one of the modelling techniques used to predict the PM_{10} concentration between dependent variable and independent variables (Dominick et al., 2012).

Sousa et al., (2007) compared Multiple linear regression (MLR) and feedforward artificial neural network (FANN) to predict the next day hourly ozone concentration using as predictors air pollutant concentrations (NO, NO_2 and O_3) and meteorological parameters (T, RH and WV). The same models, but based on principal component analysis (PCA), were also used, being referred to as principal component regression (PCR) and feedforward artificial neural network based on PC (PC-FANN), respectively. The results showed that the use of FANN led to more accurate results than linear models (MLR and PCR), due to the account of non-linearities. The application of PC in this model was considered better than using the original data, because it reduced the number of inputs and therefore decreased the model complexity. The performance indexes were similar using both approaches.

Chaloulakou et al., (2003) applied multiple linear regression method to investigate the complex relationships between meteorological and time period parameters and forecast future PM_{10} concentrations. In Athens, a study by Grivas and Chololokau (2006) used this method to predict hourly PM_{10} concentrations 24 hours in advance and the result showed that multiple regression models can be used to predict PM_{10} 24 hours in advance.

A study by Ul-Saufie et al., (2011) used Multiple Linear Regression (MLR) and Feedforward Backpropagation Artificial Neural Network (ANN) models for predicting concentration in Pulau Pinang. Multiple regression models and neural networks are examined for Seberang Jaya, Pulau Pinang with the same independent variables, enabling a comparative study of the two approaches. Model comparison statistics using Prediction Accuracy (PA), Coefficient of Determination (R^2), Index of Agreement (IA), Normalised

Absolute Error (NAE) and Root Mean Square Error (RMSE) show that ANN is the best technique compared to MLR.

Ul-Saufie et al., (2013) defined multiple regression analysis (MRA) is a popular methodology to express response of a dependent variable of several independent variables. In spite of its success, MRA presents problems in identifying the most important contributors when multicollinearity, or high correlation between the independent variables in regression equation are present. Multivariate data analysis techniques such as multiple linear regression (MLR) and Principle Component Analysis (PCA) have been proven to be effective tools to study the relationship between voluminous data such as air pollution and meteorological records. In order to indicate the performance of the models and to test the best model, five performance indicators were used to evaluate the four models (MLR, PCA-MLR, ANN, and PCA-ANN) used to predict next day, next two-day, and next three-day concentrations. The result shows that PCA-ANN is the best model to predict the next day PM_{10} concentration in Nilai, Negeri Sembilan. Meanwhile the PCA-MLR model produces the best result to predict next two-day and next three-day concentrations at the same site.

Mata (2011) compared between multiple linear regression (MLR) and artificial neural network (ANN) models for the characterization of dam behaviour under environment loads. In its lifetime, a dam can be exposed to significant water level variations and seasonal environmental temperature changes. The use of statistical models, such as multiple linear regression (MLR) models, in the analysis of a structural dam's behaviour has been well known in dam engineering since the 1950s. The two methods have the advantage of being easily implemented and of being simultaneously used, which increases the confidence in the use of these models. Finally, the results of this study reinforce the notion that statistical models are useful for establishing relations between loads and structural responses for the behaviour analysis in the safety control of concrete dams. However, ANN models showed flexibility and proved to be more adequate for months with extreme temperatures than the MLR models with the same variables.

Pires et al., (2008) compared the performance of five linear models to predict the daily average PM_{10} concentrations. The linear models implemented were multiple linear regression (MLR), principal component regression (PCR), independent component regression (ICR), quantile regression (QR), and partial least squares regression (PLSR). These models were applied to two datasets with different sizes (first set with data from three years and the second with data from six months). The results showed that the prediction of the daily mean PM_{10} concentrations was more efficient when using ICR for the smaller dataset and PLSR for the larger dataset. As presented in the referred studies, the structure definition is an important step for the model success, being followed by the optimization of its parameters. All of these studies lead to the conclusion that it is practically impossible to establish a ranking of models for predicting the daily average PM_{10} concentrations on the next day. Thus, due to the high number of variables of different nature involved in this process and the specificities of the monitoring sites, the procedure to be applied is to test simultaneously different model structures to find the one/ones with the best performances.

2.7.2 Boosted Regression Trees (BRT)

BRT is one of several techniques that aim to improve the performance of a single model by fitting many models and combining them for prediction. BRT uses two algorithms: regression trees are from the classification and regression tree (decision tree) group of models, and boosting that builds and combines a collection of models. So far, there are no research that were done by using BRT in environmental area, thus these literature reviews are more to other engineering fields.

A study by Carslaw and Taylor (2009) have developed a boosted regression trees model for hourly concentrations of nitrogen oxides (NO_X) close to a large international airport. Model development is discussed and methods to quantify model uncertainties developed. It is shown that good explanatory models can be developed and further, allowing for interactions between model variables significantly improves the model fits compared with non-interacting models. Methods are used to determine which variables exert most influence over predicted concentrations and to explore the NO_X dependency for each. Model predictions are used to estimate aircraft take-off contributions to total concentrations of NO_X and determine how these predictions are affected by annual variations in meteorological conditions and runway use patterns. Furthermore, the results relating to the aircraft contributions to total NO_X concentration are compared with those from a more detailed independent field campaign. Finally, empirical evidence that plumes from larger aircraft disperse more rapidly from the point of release compared with smaller aircraft.

In boosted regression trees (BRT) each of the individual models consists of a simple classification or regression tree, i.e. a rule-based classifier that partitions observations into groups having similar values for the response variable, based on a series of binary rules (splits) constructed from the predictor variables. The boosting algorithm uses an iterative method for developing a final model in a forward stage-wise fashion, progressively adding trees to the model, while re-weighting the data to emphasize cases poorly predicted by the previous trees. A BRT model can therefore be seen as a regression model in which each of the individual model terms is a simple regression tree (Asri et al., 2015).

Tree-based models partition the predictor space into rectangles, using a series of rules to identify regions having the most homogeneous responses to predictors. They then fit a constant to each region, with classification trees fitting the most probable class as the constant, and regression trees fitting the mean response for observations in that region, assuming normally distributed errors (Elith et al., 2008).

Razakamanarivo et al., (2011) tested and compared three different modelling techniques for various factor sets which are simple linear regression (SLM), multiple linear (MLM) models and boosted regression tree (BRT) models. Weights of the factors in the respective model were analysed for the three pool-specific models that produced the highest accuracy measurement. A regional spatial prediction of carbon stocks was performed using spatial layers derived from a digital elevation model, remote sensing imagery and expert knowledge. Results showed that BRT had the best predictive capacity for carbon stocks compared with the linear regression models.

A study by Youssef et al., (2016) used four modelling techniques, namely random forest (RF), boosted regression trees (BRT), classification and regression trees (CART), and general linear model (GLM) to produce landslide susceptibility maps and their results are compared for landslides susceptibility mapping at the Wadi Tayyah Basin, Asir Region, Saudi Arabia. Landslides locations were identified and mapped from the interpretation of different data types, including high-resolution satellite images, topographic maps, historical records, and extensive field surveys. The results revealed that the RF, BRT, CART, and GLM models produced reasonable accuracy in landslide susceptibility mapping. The outcome maps would be useful for general planned development activities in the future, such as choosing new urban areas and infrastructural activities, as well as for environmental protection.